Tags in Domain-Specific Sites - New Information?

Steinhauer, Jeremy; Delcambre, Lois M.L.; Maier, David; Lykke, Marianne; Tran, Vu H.

Published in:
JCDL'11

DOI (link to publication from Publisher):
10.1145/1998076.1998096

Publication date:
2011

Document Version
Early version, also known as pre-print

Tags in Domain-Specific Sites - New Information?

Jeremy Steinhauer,
Lois M. L. Delcambre,
David Maier
Computer Science Department
Portland State University
Portland, OR 97207-0751 USA
+1 503 725 2405
{jsteinha, lmd, maier}@cs.pdx.edu

Marianne Lykke
Dept. of Commun. & Psychology
Aalborg University
Aalborg, Denmark
+45 2125 1854
mlykke@hum.aau.dk

Vu H. Tran
Computer Science Department
Portland State University
Portland, OR 97207-0751 USA
tvu@cs.pdx.edu

ABSTRACT
If researchers use tags in retrieval applications they might assume, implicitly, that tags represent novel information, e.g., when they attribute performance improvement in their retrieval algorithm(s) to the use of tags. In this work, we investigate whether this assumption is true. We focus on the use of tags in domain-specific websites because such websites are more likely to have a coherent, discernible website structure and because the users that are searching for and tagging pages in such a site may have specific information needs (as opposed to the broad range of information needs that users have when browsing/searching the Internet at large). For this study, we assume that the application of the same tag to multiple pages provides an indication that those pages are related. To determine whether this indication of relatedness is contributing new information, we first measure whether pages with common tag(s) could have been deemed as related based on site structure as measured by shortest navigational distance between pages. Second, we measure whether or not tags could have been determined algorithmically based on standard tf-idf scores of terms on the page. Based on our analysis of two different sites, we found that tags contribute novel information that is not discernible from site structure or site/page content.

Categories and Subject Descriptors
H.3.3[**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms
Human Factors, Measurement, Verification

Keywords
Tags

1. INTRODUCTION
The Web 2.0 movement emphasizes the value of allowing end-users to develop and augment content, e.g., by applying tags. A user often tags items of content, i.e., applies a term or phrase to categorize those items, in order to find them easily at some later time. Since tags applied by one user are typically visible to all users, the set of tags applied by a community of users often results in a more robust description of content than would have been generated by any one individual [1]. It has been shown that tags closely reflect the vocabulary that users use in searches [2] and that tagging may be a scalable way to get descriptive information about documents, compared to metadata provided by authors and indexers [3].

Tags are not without their problems. Tags are often 'wrong.' They may not accurately reflect the content to which they are applied; they may contain misspellings, nonsensical words, or profanities [3]. Tags may contain unusual characters. For example, delicious[1] only allows single word tags but we have observed people using: *breast_cancer*, *breast.cancer*, or *breastcancer* to represent the phrase *breast cancer* when they tag. Also, Munk et al. showed that users tend to tag with general terms such as *disease* or *cancer* rather more specific term like *stage II colorectal cancer* [3]. However, even with all of these problems, it has been shown that, as more tags are added, the tagging vocabulary for a collection tends to converge on a set of frequently used terms often referred to as a folksonomy [3].

Tagging facilities are often provided on sites that are essentially (just) a collection of objects (e.g., books in LibraryThing[2], photos in flickr[3], or research papers in CiteULike[4]). Such sites typically have no internal structure (i.e., no links) and they tend to rely on tag searching and browsing to find information on the sites.

In contrast, our interest is in tagging facilities for domain-specific websites where, often, considerable work has gone into organizing information into a useful browse structure, as reflected in the site structure. We are collaborating with the Danish Cancer Society where we have the unique opportunity[5] to study tags and tagging behavior on a domain-specific site over an extended period of time.

The research presented here is a preliminary investigation into tags in two, domain-specific sites to determine whether tags contain information beyond the information contained in the site structure and page content. We are interested in site structure because pages that are directly linked are likely to be related and, more generally, pages that are close together may be more related than pages that are far apart. We are interested in page content

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.

[1] www.delicious.com

[2] http://www.librarything.com/

[3] http://www.flickr.com/

[4] http://www.citeulike.org/

[5] This tagging facility is expected to be implemented in cancer.dk in the Summer of 2011.

because some studies have shown that users often tag pages with words or phrases that appear on the page.

The contribution of this paper is the detailed analysis of the delicious tags for cancer.org (the website for the American Cancer Society) and simplyrecipes.com (a website to share recipes). We investigate whether pages linked by shared tags could be determined by an analysis of shortest distances between those pages. We also investigate whether tag terms could be determined from an analysis of term frequency data from the pages on which they are applied.

We examine related work, show how we gathered data, analyze the data and discuss our results in Sections 2, 3, 4, and 5 repectively.

2. RELATED WORK

Much work has gone into studying how tags can be incorporated into various information retrieval techniques [4-8].

The work that most closely resembles the work we describe here is that of Heymann and Garcia-Molina [9] whose research compares tags with a controlled vocabulary. They compared book tagging data from LibraryThing with the Library of Congress Subject Headings, LCSH, for the same books. They found that 48% of the words in the controlled vocabulary had an exact or nearly exact equivalent in the tags. When comparing tags and terms that had similar semantic meaning, based on an analysis of Wikipedia[6] pages, they found that most tags could be mapped semantically to LCSH terms. However, when looking at whether semantically equivalent terms were being applied to the same books they found that only 21% of terms matched in all cases and 56% had at least one book in common. So they concluded that taggers and professional indexers mark books with semantically similar terms but differ in their application of those terms.

Another closely related study by Marshall [10] examined how tags compared to other forms of user-generated metadata. In this study she gathered a set of similar images from flickr. She then compared the tag information for these photos to the other user-generated information about the photo, the title, and the narrative caption. She found that tags tended to contain a subset of the terms used in the caption and titles even when common stop words were removed. Stop words are words that are assumed to contain little or no significance for retrieval such as *the* and *or*. It is common in information retrieval to ignore stop words. She concluded that narrative captions for photos provide a more robust description than tags and could be used in much the same ways as tags.

Golub et al. [11] looked at whether the quality of tags could be enhanced by offering suggestions from a controlled vocabulary. They created a system that allowed for free tagging but would also suggest controlled vocabulary terms based on the title of the document to be tagged. While they found that users did not like their user interface but when they used it the users tagged with more facets then when they just free tagged.

3. DATA GATHERING

delicious is a public online bookmarking service that allows users to save URLs and apply tags to them. In the Spring of 2010, we collected all of the tag-URL applications made in delicious during a one-week period, using their standard API. We analyzed the sites that were tagged during this period to find sites that were domain-specific, rich in information, and well structured with a reasonable number of tags. We chose to investigate cancer.org and simplyrecipes.com.

3.1 Crawling

We crawled our selected sites using Nutch, an open source web crawler developed by the Apache Software Foundation. The crawls of cancer.org and simplyrecipes.com retrieved 8591 and 1945 unique pages respectively.

We extracted document-to-document links to generate a directed graph of the site structure where each directed edge represented a link from one page to another. We applied the Floyd-Warshall shortest path algorithm to this graph to compute the shortest path between every pair of pages on the site.

We used the same crawl to extract term frequency information with which we calculated tf-idf scores for each unique term on each page. Tf-idf, term frequency multiplied by inverse document frequency, is a score often used to determine a term's importance in a document relative to the entire collection. Tf is the frequency of a term t in a document normalized to all the terms in that document d

$$\text{tf} = \frac{t}{d}$$

and idf is log of the inverse of the number of documents in the collection that contain term n, normalized to the total number of documents in the collection m

$$\text{idf} = \log\frac{m}{n}$$

and tf-idf is just tf*idf.

While verifying the data from our crawls, we noticed a set of URLs with a similar structure in cancer.org that mostly returned *page not found* errors plus a few redirects to new pages. Since the URLs in this set were linked to from other pages in cancer.org, we were suspicious that they may have been previously valid URLs. We looked these URLs up in the Wayback Machine[7], a website that archives frequent snapshots of the web and allows you to search for a page based on date. We found that all of the pages in this set ceased to exist at about the same time in what appeared to have been a site redesign.

3.2 Tags from delicious

We created a tool to extract tag-URL data from the delicious web interface for every URL found by our crawl[8]. We found 1032 unique tags, 2202 tag instances, and 117 tagged pages for cancer.org and 1945 pages, 6672 unique tags, 34337 tag instances, and 1168 tagged pages for simply recipes.

As noted above, delicious users circumvent the single-word tagging requirement by making phrases without spaces such as *breast_cancer*, *breast.cancer*, and *breastcancer*. We developed an algorithm to detect these types of phrases in tags and to determine where they appear in the document collection. First, we deemed non-alphabetic characters in the middle of a tag to be word delimiters and split the tag into a phrase accordingly. For each remaining tag, we found all of the terms from the document collection (extracted from our crawls) that matched a prefix of the

[6] www.wikipedia.com

[7] http://www.archive.org/web/web.php

[8] This process was slowed because after every 50 requests to delicious our connection was refused for 2 hours. We tried a couple of different throttling mechanisms but were unable to circumvent this limitation.

tag. Then we in turn removed each of those terms from the head of the tag and recursively searched for a prefix using the remaining string. If, through this method we found terms in the collection that completely matched the tag, we searched the term indexes from our crawl to see if those terms ever appear consecutively within the documents in our collection.

Another issue that we discovered during our verification process was that Nutch and delicious handle URLs that redirect differently. Nutch creates a hash of the content of every URL it crawls and when the hash is the same, whether through a URL redirect or an alias for a URL, it indexes only the first one it found and points the rest of the URLs to that first one. Delicious handles aliases in the same way; it keeps a record of both pages but the tag data for the two (aliased) pages is the same. However, for a URL that results in a redirect, delicious creates a separate tag list even though the URL may redirect to a page that already has tags. So we identified the URLs that resulted in a redirect, gathered tags for those URLs, and then merged those tags with the tags for the (actual) page that these URLs redirected to.

Table 1 Distribution of pairs of pages of a certain shortest distance apart shown as percentages.

	Cancer.org								Simplyrecipes.com							
	Pairs of pages		Pairs of pages that share tag(s)						Pairs of Pages		Pairs of pages that share tag(s)					
dist.	All	Tagged	Any	1	2	3	4	5	All	Tagged	Any	1	2	3	4	5
1	0.5	6.7	11.1	10.0	14.4	21.1	19.4	16.6	2.8	3.9	3.8	5.3	3.4	3.6	3.3	3.0
2	5.6	18.8	20.3	17.8	30.2	33.3	30.5	50.0	44.1	58.8	59.3	54.4	56.1	58.9	62.3	64.1
3	44.6	48.9	26.7	27.5	25.0	16.6	22.2	27.7	45.0	34.6	34.2	36.8	37.4	34.9	31.9	30.7
4	37.9	19.5	23.6	26.2	12.6	7.78	11.1	5.5	7.6	2.4	2.6	3.4	2.9	2.4	2.3	2.1
5	6.4	4.9	11.2	11.6	9.74	11.1	8.3	16.6	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
total	8591^2	117^2	3185	2631	380	90	36	18	1945^2	1168^2	1.2M	269k	288k	237k	166k	104k

4. ANALYSIS AND RESULTS

We report on the types of analysis we did for cancer.org and simplerecipes.com and the results of these analyses in this section.

4.1 Shortest distance analysis

To determine whether relationships created between pages that share tags could be determined from the site structure, we chose to analyze the shortest distance between pairs of pages. Since the sites we chose appear to be well-structured, we expected that as pages share more tags, they would be closer together based on the shortest distance between pages.

Table 1 shows, for both cancer.org and simplyrecipes.com (on the left and right side of the figure, respectively), the percentages of pairs of pages that are of a certain distance apart and the total number of pairs of pages that these percentages are based on. This data is broken out by all pages and tagged pages. Then for tagged pages, the tables shows the percentages of pages that share at least one tag (in column labeled "Any" and the percentages of pages that share 1, 2, …, 5 tags.

For both sites, when we look at the shortest distance between pages that share any (the any column in Table 1) tags and all (Table 1, all column) pairs of pages we see that pages that share tags tend to be closer together. We also see clearly in simplyrecipes.com that as pages share more tags, the average shortest distance between pages decreases. This effect is evident in cancer.org except for when pages have 4 tags in common where we see a slight increase in the distance apart; however, the number of pairs of pages that share 4 tags is relatively low (36).

Our main research question is: is the closeness of pages in the site structure enough to let us detect the relationship that is indicated when two pages share tag(s)? For simplyrecipes.com we see that the site is highly interconnected (just under half of the pages are 2 clicks or less apart. Thus it would not be useful to infer any sort of relationship between pages farther than one click apart since each page would be deemed to be related to half the site.

We found that cancer.org is a deeper site that is less interconnected. Still, the average number of pages within a distance of two clicks from any given page is 481 (.056 * 8591) where the chance that a page shares a tag with one of them is quite low ($646/8591^2$ = .0001). So we see that a shortest distance analysis of site structure is unable to detect the relationships between pages that are of distance 2 or greater away for either site. So the relatedness of 89.9% and 96.2% pairs of pages that share one or more tags (i.e., the percentage of pages that share one or more tags that are farther than 1 click away) is new information for cancer.org and simplyrecipes.com, respectively, compared to the relatedness of pages based on site structure as measured by shortest distance.

During our analysis, we noticed that every page in both cancer.org and simplyrecipes.com had a link back to the homepage (as well as other standard types of links). In order to see whether the link to the homepage was causing pages to be closer together than they would be based on the remaining site structure, we recalculated shortest distances between pages where paths that included the homepage were eliminated. We found the results from these new shortest distances were not significantly different from those shown in Table 1. This result may be due to the fact that many of the other standard links on the pages linked to pages one click away from the homepage.

For cancer.org, we discovered that while the top level directory *espanol* contained 28% of the cancer.org overall content it contained less than 4% of the tagged pages. We assume that this distribution is a reflection of the population of delicious users, e.g. few Spanish speaking users, rather than a judgment of the value of the content within this section.

This nonuniform distribution of tagged data means that pairs of tagged pages are closer together regardless of whether they share a tag. We can see this property clearly in Table 1 when we compare all (Table 1, all column) pairs of pages to (Table 1, tagged column) pairs of tagged pages.

4.2 Term frequency analysis

We are interested in whether tag data could have been generated based on an analysis of tem frequency data. For this analysis we first checked whether tag terms were found on the pages to which they are applied. When they were, we used tf-idf scores to determine the terms importance to the page.

The first part of Table 2 shows how much tag terms overlap with page terms. We see that for cancer.org 74% of tag terms do not appear on the pages they are applied to. When we added phrases using the method described in Section 3.2, 67% of the tags did not appear on the page to which they were applied. For

simplyrecipes.com we see that 43% of the tags (and 26% of tags, when phrases were considered) do not appear on the pages they are applied. When there is no overlap between page and tag terms it means that the tags could not have been determined by page content alone and therefore represents wholly new information.

In the second part of Table 2, we report the percentage of tag terms whose tf-idf scores fall within the given range of tf-idf scores for terms on that page. We only used single term tags for this analysis as calculating tf-idf scores for phrases of unlimited length is challenging. Here we see that the majority of tag terms have a lower than average tf-idf score. We focused on terms that were two standard deviations above the mean tf-idf score because we were looking for scores that were significantly higher than the scores for other terms on the page. We see that 96.7% (74.91+.8371*26.09) and 89.6% (42.96 +.8173*57.04) of all tagging data does not fall above this level for cancer.org and simply recipes.com, respectively, and therefore it is extremely unlikely that we could determine tags from page content alone using tf-idf.

Table 2 Percentages of tag and page term frequency overlap and relative tf-idf score distribution for those terms

	Cancer.org	Simplyrecipes.com
Tag Found	26.09%	57.04%
Phrase/Tag Found	33.36%	74.31%
Below Average TF-IDF	53.98%	58.12%
Above Average TF-IDF	46.02%	41.88%
Above 1 std. dev.	27.27%	23.66%
Above 2 std. dev.	16.29%	18.27%
Highest TF-IDF	6.25%	4.00%

5. CONCLUSIONS AND FUTURE WORK

While our results provide evidence that tags contain information beyond that inherent in site structure and page content, cancer.org had only 117 pages of 8591 pages tagged. Certainly the cancer.org site redesign we detected, that occurred about 2 years ago, may have contributed to this sparseness of the data. Delicious has been collecting tags since 2003; for pages that changed or moved in the redesign, we lost over 5 years of data. We identified about 400 lost pages based on broken links alone. For those lost pages we were able to find about 100 tag instances associated with them. Simplyrecipes.com had many more tags with well over half its pages tagged. Since both site produced similar results, we can surmise that with more tagging data, the results for cancer.org would likely be similar to the results shown here.

When we examined terms used for tagging. Most of the top tags used on the site were quite generic. For example, *cancer*, *health*, *medical*, and *acs* (American Cancer Society) were the top four tags used on the cancer.org site. The predominance of these types of tags was likely influenced by the source of the tagging data. Delicious bookmarks are not limited to one domain. Users can group their domains of interest using these types of generic terms. We hypothesize that tags applied in a domain-specific site will likely contain fewer generic tags since such tags would apply to the entire site.

We recognize that shortest path is not the only way to use site structure to try to determine relatedness of pages. We are interested in whether an analysis of path density between pages (the number of unique paths between two pages) may be higher when the pages share tags.

We are also interested in expanding our analysis using tf-idf scores by examining the distance between tf-idf-weighted term vectors for all pages. It is possible that the relationship between pages that share the same tags could be determined based by on this analysis.

6. ACKNOWLEDGEMENTS

We gratefully acknowledge: Tyler Hayes who helped develop an early crawler and gathered initial data for this research; Marit Kristine Ådland who is part of our larger research project investigating tagging in domain-specific collections; and Sarann Bielavitz and Scott Britell for their input on this project. We also gratefully acknowledge our collaboration with colleagues at the Danish National Cancer Society.

This work was supported in part by NSF grant number 0812260. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] Golder, S., Huberman, B. 2006. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science* 32(2).

[2] Carman, M., Baillie, M., Gwadera, R., Crestani, F. 2009. A Statistical Comparison of Tag and Query Logs. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[3] Munk, T., Mork, K. 2007. Folksonomy, The Power Law & The Significance of the Least Effort. *Knowledge Organization* 34(1).

[4] Ramage, D., Heymann, P., Manning, C., Garcia-Molina, H. 2008. Clustering the Tagged Web. *Second ACM International Conference on Web Search and Data Mining*.

[5] Begelman, G., Keller, P., Smadja, F. 2006. Automated Tag Clustering: Improving Search and Exploration in the Tag Space. *www2006.*

[6] Bao, S., Wu, X., Fei, B., Xue, G., Su, Z., Yu, Y. 2007. Optimizing Web Search Using Social Annotations. *www2007.*

[7] Liang, H., Xu, Y., Li, Y., Nayak, R. 2008. Collaborative Filtering Recommender Systems Using Tag Information. *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*.

[8] Nakamoto, R., Nakajima, S., Miyazaki, J., Uemura, S., Kato, H., Inagaki, Y. 2008. Reasonable Tag-Based Collaborative Filtering For Social Tagging Systems. *Proceeding of the 2nd ACM workshop on Information Credibility on the Web*.

[9] Heymann, P., Garcia-Molina, H. 2008. Contrasting Controlled Vocabulary and Tagging. *Second ACM International Conferenceon on Web Search and Data Mining.*

[10] Marshall, C. 2009. No Bull, No Spin: A Comparison of Tags with Other Forms of User Metadata. *Proceedings of the 2009 Joint International Conference on Digital Libraries*, 241-250.

[11] Golub, K., Jones, C., Lykke Nielsen, M., Matthews, B., Moon, J., Puzoń, B., Tudhope, D. 2009. EnTag: Enhancing Social Tagging for Discovery. *Joint Conference on Digital Libraries.*